Knowledge Management through a Fully Extensible, Schema Independent, XML Database

H. Direen[1], C. Brandin[1], M. Jones[1], C. Hedgepeth[2],D. Shin[2]

[1]NeoCore Inc., Colorado Springs, CO, USA, www.neocore.com

[2]Varro Technologies Inc., Philidelphia, PA, USA, www.varrotech.com

Abstract – Information in any field, including bioinformatics, is a combination of data and the data's context. A system that manages information must be able to handle both aspects of information: context and data. The problem with most systems (databases in particular) is that the context must be predefined. In a field that is developing as fast as bioinformatics, it is as impossible to predefine all of the context as it is to predefine all of the data that is being generated. A knowledge management system must handle context or metadata as freely and flexibly as it handles data. NeoCore has developed a fully extensible, schema independent, XML-based information management system that meets this requirement.

Keywords – Extensible Database, Schema Independent, Knowledge Management, Information Management, XIMS, XML, Bioinformatics, NeoCore, Varro

I. Introduction

THE intent of storing data in genomic databases, proteomic databases, and many other databases in use, goes well beyond creating repositories of data. The primary intent of these data repositories is the building, management, and access of knowledge. Knowledge is understanding (information) gained through experience or study. A knowledge management system is a system that allows one to capture, store, query and disseminate information which is gained through experience or study. A key problem with current data repositories is that the schema, which includes the data structure and the type of data stored, must be predefined. This implies that the creator of the repository (database) must have *a priori* knowledge of all information being stored in the database, which precludes handling "information gained through experience or study".

Information is the combination of data and context. If I give you the sequence of letters 'seat', without any context, you have no way of knowing what I am referring to. I could be referring to an uncomfortable seat on the plane ride to this conference, or I could be referring to a short amino acid sequence contained within the Aspartokinase I protein in E-Coli. The point is that data without context is incomplete. It is only when we combine data with its context that we have information. A system that locks context and only allows data to be modified, which is typical of traditional database systems, is unsuitable for a knowledge management system. Ideally, a knowledge management system will handle context (metadata) as freely as it handles data.

The eXtensible Markup Language, XML, is well suited for knowledge management systems [3] in that XML, by design, pairs data with its context via the hierarchical tag structure. XML is the first widely adopted standard for expressing information as opposed to just data. This characteristic has precipitated the acceptance of XML as the preferred transport mechanism for data exchange. XML is also rapidly becoming an accepted data language for bioinformatics [1]. There are a number of groups developing standards in this area, including: the Bioinformatic Sequence Markup Language (BSML) (http:// www. visualgenomics. com/ products/ index.html); the BIOpolymer Markup Language (BioML) (http:// www. proteometrics. com/ BIOML/); GAME (http:// www. bioxml. org/ Projects/ game); plus many others. Paul Gordon has a website devoted to XML for Molecular Biology: (http:// maggie.cbr.nrc.ca/ ~gordonp/ xml/). The unprecedented degree of flexibility and extensibility of XML in terms of its ability to capture information is what makes it ideal for knowledge management and for use in bioinformatics.

Molecular biology, genomics, and proteomics are rapidly developing fields. The mapping of the entire human genome, completed this year, represents a major achievement in the history of mankind. Yet this achievement is only the tip of the ice-berg in terms of the knowledge that will be garnered in the years to come as we begin to unravel the information contained within the human genome, the genomes of all living creatures, and the now exploding field of proteomics. Who would be willing, at this juncture, to predefine all of the contextual fields that will result from these areas of research so that we might capture all of the information generated? In a knowledge management system, it is as impossible to predefine all of the context fields as it is to predefine all of the data. Yet this is the approach of traditional database systems. What is required is a system that treats context (metadata) as flexibly as it treats data. XML provides a basis for treating data and metadata. This paper describes an XML-based information managememt system, created by NeoCore, that handles metadata in exactly the same fashion as it handles data. The example background will be drawn from bioinformatics, but the NeoCore XIMS (XML Information Management System) meets broader requirements [5].

II. NeoCore XML Information Management System

Traditional databases were designed to manage data by creating a static framework to contain dynamic data. That is, data elements can be managed as long as all metadata

0-7803-7211-5/01$10.00©2001 IEEE

Report Documentation Page

Report Date 25 Oct 2001	**Report Type** N/A	**Dates Covered (from... to)** -

Title and Subtitle
Knowledge Management Through a Fully Extensible, Schema Independent, XML Database

Contract Number

Grant Number

Program Element Number

Author(s)

Project Number

Task Number

Work Unit Number

Performing Organization Name(s) and Address(es)
NeoCore Inc Colorado Springs, CO

Performing Organization Report Number

Sponsoring/Monitoring Agency Name(s) and Address(es)
US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500

Sponsor/Monitor's Acronym(s)

Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.

Abstract

Subject Terms

Report Classification
unclassified

Classification of this page
unclassified

Classification of Abstract
unclassified

Limitation of Abstract
UU

Number of Pages
4

(all the data's context) has been established in advance. This methodology falls short, by half, of true information or knowledge management. The solution to this dilemma is to manage metadata in exactly the same dynamic way as the data component. This offers two significant (if not profound) advantages. First, constraints on the use of dynamic data types are removed. New data types may simply be defined and added to the database. Second, the processes of database design and configuration (usually the most labor and time consuming aspect of system design) are reduced to practically nothing. The NeoCore XIMS (XML Information Management System) [1] has been designed with precisely this feature: metadata and data are handled in the same dynamic way. The XIMS was designed to achieve the following goals:
1. *Dynamic management of metadata.* All information is represented in an internal format called "Information Couplets" – pairings of data and complete metadata. Information couplets are treated as patterns, and those patterns can be arbitrary. Consequently, there are no rows or columns, and there is no need to predefine indices. Patterns can be as common or as unique as desired. A piece of information can be associated with a single fragment of information without having to be predefined, pre-allocated, or added to another information fragment. This means that an entirely new data type can be added to an application at any time without having to do anything to the database at all.
2. *Immediate availability of information.* Information is "indexed" as soon as it is posted. Actually, there is no separate indexing process because there is no need to specify what should be indexed. The NeoCore XIMS automatically generates data patterns that allow information to be retrieved based on fully or partially qualified queries. Complex queries are accommodated through a hierarchical vector convergence algorithm, which quickly converges sets of individual pattern matches to locate information fragments based on multiple criteria. The vector convergence process operates on any information that has been posted, requiring no predefinition of any sort. The structure of the indices created within the XIMS gives flat access time to all nodes within the system. There are no access time penalties based on the structure of the data stored in the XIMS.
3. *Schema independence.* The NeoCore XIMS was designed to be oblivious to schemas or DTDs (XML Document Type Definitions). This is a more important feature than it may seem at first blush. Most XML data management systems require that all XML information be described by a schema or DTD for data mapping reasons. The problem schema dependence imposes is that it destroys the ability to have heterogeneous data within similar document types. For example, suppose you want to add a new field to some, but not all, documents of the same type. How will the database know which schema applies unless a new schema is supplied? How will it know whether all the other documents of the same type need to be changed, or whether they should be treated as different document types? How will other applications know about this new document type? Ultimately, schema dependence makes it virtually impossible to use XML's most attractive feature – its extensibility. Schema independence implies that there is no database design that needs to be done, and no penalty is imposed for change.
4. *Scalability.* The XIMS was designed to manage huge repositories (terabytes) of XML documents of all types. This was achieved by treating XML documents as aggregations of information. The NeoCore XIMS is aware of, but functionally oblivious to, the document-centric structure of XML information. XML data management systems are notorious for not scaling well – both as document size increases, and as the number of documents increases. The NeoCore XIMS was designed to seamlessly scale to very large information management requirements, exhibiting remarkably flat performance as system size increases. Individual document size has no bearing on performance.
5. *Efficient use of storage.* Information Couplets represent a fundamental means of storing and managing information. Breaking the information into couplets along with efficient indexing using NeoCore's patented Digital Pattern Processing (DPP) technology [2], creates a very efficient storage format. The upshot is that the amount of storage used, for everything combined, adds up to between 1x and 2x the size of the XML documents alone. This includes the documents, all indices, access control information – everything. This compares very favorably with all other methods of managing information, requiring less than half the storage of any database management system, and a tiny fraction of the space used by DOMs (XML Document Object Models).
6. *Speed.* The NeoCore XIMS was designed from the onset to be very fast. This was largely achieved through the use of NeoCore's DPP technology, information couplets, and other internal set convergence routines. By focusing on the nature of XML, instead of prior methods of managing data, NeoCore was able to tailor their XIMS precisely to the challenges posed by the nature of self describing, heterogeneous information – a very different paradigm than existed when older data management methods were developed. The result is a software system that outruns, by a wide margin, any networks' ability to transport data. A relational database management system, for example, can process in the range of 1-10 XML transactions per second. Typical "Native" XML databases can process, in the best-case scenario, about 25 XML transactions per second. A fast Ethernet connection saturates at about 1000 simple XML transactions per second. The NeoCore XIMS can process tens-of-thousands of XML transactions per second, meaning that it will be able to keep up with any networks' ability to process transactions – now, and in the future.

The features noted above make the NeoCore XIMS a very powerful knowledge management system. The next

[1] The following patents protect NeoCore technology: US Patent #5,742,611 (April 21, 1998), US Patent #5,942,002 (August 24, 1999), US Patent #6,157,617 (December 5, 2000), US Patent #6,167,400 (December 26, 2000), Other U.S. and international patents pending.

section elucidates these features through examples from the bioinformatics arena.

III. Information Management in Bioinformatics

With the vast amount of biological information stored in databases throughout the world, the life scientists of today spend a tremendous amount of their time gathering, sorting, manipulating, and adding to this vast pool of knowledge. The information contained in these disparate databases is extremely heterogenous and stored in widely varying formats. This makes finding, querying, and working with requisite information pertinent to the individual researcher an ominous task. Varro Technologies is in the process of building a distributed, knowledge management network, Varrogen, built on the NeoCore XIMS technology. Varrogen will give researchers a single point of contact for the disparate, public and private databases distributed throughout the world. In addition to being a single point of contact, the researcher or research organization may contract with Varro to store and share (in a controlled environment) the information they are garnering, creating, and building. The following example will highlight some of the issues in biological information management and solutions to those issues.

Suppose our researcher is studying cystic fibrosis (CF). There is a plethora of information on various aspects of CF stored in both public and private databases throughout the world. A researcher's first task is to find and gather all related information pertinent to her specific area of interest. The researcher may start at the public databases NCBI, EMLB, or DDBJ depending on location, or the researcher may start with Varrogen. Specific gene information may be downloaded in a flat file format, or alternatively it may be downloaded in XML format. Flat file format is a text file that may be stored on the computer. The problem is, once a large number of files have been gathered, there is no easy way to later query and find specific information gathered. Nor is there an easy, clean way to later add to the information. With Varrogen the information gathered will automatically be put into XML format. Also, since XML has become a standard for data transport, most databases give this option for download. The XML information may be directly imported into a NeoCore XIMS without concern of predefining what information is being stored (a schema for the data is not required). The server understands XML and the XML structure. The information is stored and fully indexed automatically. There is no need to predefine key fields for later access of the information store. As different types of information about CF are gathered from various sources, the information may simply be added to the XIMS without regard to predefining a database structure. An example of a piece of information garnered from the NCBI genebank, in XML format, is:

```
<genbank_entry>
    <header>
        <locus  bp=“741” type=“mRNA” ></locus>
        <definition>
            Homo sapiens cystic fibrosis transmembrane
            conductance regulator...
        </definition>
        <accession>NM_000492</accession>
        <version>NM_000492.2 GI:6995995</version>
    </header>
    <Comment>
        This record has been curated by NCBI...
        The protein encoded by this gene is a member
        of the super family of ATP-binding cassette...
    </Comment>
     <features>
       <feature chromosome=“7” ></feature>
       <feature map=“7q32.1” ></feature>
       <feature gene=“CFTR” ></feature>
       <feature product=“ABC_membrane” ></feature>
       <feature note=“ABC transporter” ></feature>
       <feature note=“ATP-binding” ></feature>
     </features>
    <base_count a=”1886” c=”1181” ...  ></base_count>
    <origin>
        aattggaagcaaatgacatcacagcaggtc...
    </origin>
</genbank_entry>
```

It is not important to understand the XML format other than to notice that there is a hierarchical structure and that all the data is encased within descriptor tags. These tags are the context or metadata of the data items. Using the NeoCore XIMS or the Varrogen system, the researcher could easily query for something like, “find all ATP-binding, ABC transporter genes on chromosome 7 and return all comment fields contained within the gene information found”.

The researcher may be involved in studying transport pathways that the CFTR gene is involved in. For instance, she may find a regulatory mechanism that ties the sodium ion absorption rates effected by CFTR with outwardly rectifying chloride channels. The researcher will want to add the results from her findings to certain documents or records in the database. New information may be added to the database by simply defining the tag structure for the new data types, querying the database to find the specific record, and inserting the new information (data plus context). The various bioinformatic standards groups may be helpful in defining standard tag names and structure for the new information. The new information may be added without affecting other records (or documents) within the database in any way. In standard database technology, to add a new data type, a new field must be defined and that field becomes a column in the database. This changes the database structure and adds cell locations in all records of the same type, even if they are not being used by the majority of records. The database must be brought down and rebuilt with the new field before the new data item can be entered. With the NeoCore XIMS, the new information (data plus tag) is simply added to the record of interest. This new information is automatically indexed and immediately available. No changes to the database structure are required. The NeoCore XIMS handles metadata in the same fashion and as flexibly as it handles data. As it is discovered, knowledge can be added to the XIMS.

The Varrogen system will allow a group of collaborators to share a common information pool. As an example, suppose a clinical professor in the gastroenterology division of a research hospital receives a grant to identify new mutations in genes involved with colorectal cancer. He is studying genes identified mainly from patient samples in his own clinical practice. He spends 70% of his time in the lab and 30% in the clinic. Because of his hospital and teaching responsibilities, our professor can only focus on a few priorities in the lab. His clinic is thriving so he has many more clinical samples than he could ever possibly pursue. He is filling freezer after freezer with samples and only has the time to do some basic sequence analysis to determine if they are related to his primary grant-related focus. Using the Varro system, the researcher can:

1. Become a member of the Varro distributed network.
2. Establish an on-line library of the clinical samples on which he is not working, along with the basic, screening sequencing results that he has done.
3. Choose a setting that allows him to grant selective access to his data (that is, pending his approval).

Other researchers interested in colorectal cancer may establish a collaboration with the professor in order to obtain the excess clinical samples and to share research information. The professor may define a network consisting of himself and the collaborators that he approves. The network will allow the group to share and manipulate data in a common environment. The users of the neworked information system may access information that is stored, add new information, define new data types, and change information. Each collaborator may add or define new data types such as histological data, genotyping data, patient history data, and whatever types of data become apparent during the course of research. These new information types can be added without effecting the other users. This information is automatically indexed and immediately available to all users without changing the structure of the database. The collaborators can build a shared knowledge base which may be key to discovering cures to colorectal cancer.

IV. Conclusions

With information being defined as context plus data, and knowledge defined as an understanding gained through experience or study, it is clear that a knowledge management system must be able to handle both data and context in a flexible and symmetric way. XML is a standard that is becoming widely accepted in the bioinformatics community. A primary reason for its acceptance is that it provides a hierarchical, flexible structure for capturing the diverse, heterogenous information being generated in the exploding fields of microbiology, including genomics and proteomics. There are a number of groups and organizations that are developing and defining the ontologies and taxonomies pertinent in the bioinformatics fields in terms of the XML format. NeoCore has developed an XML-based information management system that uniquely handles the components of information, both context and data, in a symmetric and dynamic fashion. The server is schema independent, so that there is no up-front database design. Information may be added to the server without predefining fields of any sort. All information added is completely and automatically indexed. There is no reason to predefine key fields to index for later access and retrieval of information stored in the database. All information in the database may be accessed via queries that contain full or partial context information, and queries may even be data only where the goal is to discover context. New information may be added to specific documents or records in the database without affecting in anyway other similar records in the database. The new information does not create unused fields in all other records of similar type. The NeoCore XIMS is very efficient in storing the XML information. The amount of storage used for both the information and all indexing is only one to two times the amount of storage required to simply store the original XML document. The NeoCore XIMS is extremely fast. The server, on typical systems, can process tens-of-thousands of XML transactions per second, running much faster than the 1000 or so simple XML transactions per second that saturates a fast Ethernet connection. These characteristics make the NeoCore XIMS a very powerful knowledge management system.

Varro Technologies is building a distributed knowledge management system, Varrogen, based on the NeoCore XIMS, which will give researchers a single point of contact for the disparate, public and private databases distributed throughout the world. Researchers will have comprehensive access to procure (when necessary) and share all genomic related information from proprietary content partners, public databases, and individual researchers.

References

[1] F. Achard, G. Vayssiex, and E. Barilliot, "XML, bioinformatics and data integration", Bioinformatics Review, Oxford Univ. Press, Vol. 17 no. 2, pp. 115-125, 2001

[2] C. Brandin, "A Definition of Digital Pattern Processing", NeoCore Technical Paper, Feb. 2001.

[3] J. Cook, "XML Sets Stage for Efficient Knowledge Management", IT Pro, pp. 55-57, June 2000.

[4] K.H.Cheung, and D.G. Shin, "A Graph-Based Meta-Data Framework for Interoperation between Genome Databases", Proceedings IEEE BIBE, pp. 109-117, 2000.

[5] J. Thompson, "Mission Impossible XML (MI-XML)", XML Journal, vol. 2, issue 3, 2001.

[6] R.K. Wong, F. Lam, S. Grahm, and W. Shui, "An XML Repository for Molecular Sequence Data", Proceedings IEEE BIBE, pp. 35-42, 2000.